Data enriched linear regression

Aiyou Chen
Google Inc.

Art B. Owen*
Stanford University

Minghui Shi
Google Inc.

March 2013

Abstract

We present a linear regression method for predictions on a small data set making use of a second possibly biased data set that may be much larger. Our method fits linear regressions to the two data sets while penalizing the difference between predictions made by those two models. The resulting algorithm is a shrinkage method similar to those used in small area estimation. Our main result is a Stein-type finding for Gaussian responses: when the model has 5 or more coefficients and 10 or more error degrees of freedom, it becomes inadmissible to use only the small data set, no matter how large the bias is. We also present both plug-in and AICc-based methods to tune the penalty parameter. Most of our results use an L_2 penalty, but we also obtain formulas for L_1 penalized estimates when the model is specialized to the location setting.

Keywords: Data fusion, Small area estimation, Stein estimation, Transfer learning

1 Introduction

The problem we consider here is how to combine linear regressions based on data from two sources. There is a small data set of expensive high quality observations and a possibly much larger data set with less costly observations. The big data set is thought to have similar but not identical statistical characteristics to the small one. The conditional expectation might be different there or the predictor variables might have been measured in somewhat different ways. The motivating application comes from within Google. The small data set is a panel of consumers, selected by a probability sample, who are paid to share their internet viewing data along with other data on television viewing. There is a second and potentially much larger panel, not selected by a probability sample who have opted in to the data collection process.

The goal is to make predictions for the population from which the smaller sample was drawn. If the data are identically distributed in both samples, we

*Art Owen was a paid consultant for this project; it was not part of his Stanford responsibilities.

should simply pool them. If the big data set is completely different from the small one, then it makes sense to ignore it and fit only to the smaller data set.

Many settings are intermediate between these extremes: the big data set is similar but not necessarily identical to the small one. We stand to benefit from using the big data set at the risk of introducing some bias. Our goal is to glean some information from the larger data set to increase accuracy for the smaller one. The difficulty is that our best information about how the two populations are similar is our samples from them.

The motivating problem at Google has some differences from the problem we consider here. There were response variables observed in the small sample that were not observed in the large one and the goal was to study the joint distribution of those responses. That problem also had binary responses instead of the continuous ones considered here. This paper studies linear regression because it is more amenable to theoretical analysis and thus allows us to explain the results we saw.

The linear regression method we use is a hybrid between simply pooling the two data sets and fitting separate models to them. As explained in more detail below, we apply shrinkage methods penalizing the difference between the regression coefficients for the two data sets. Both the specific penalties we use, and our tuning strategies, reflect our greater interest in the small data set. Our goal is to enrich the analysis of the smaller data set using possibly biased data from the larger one.

Section 2 presents our notation and introduces L_1 and L_2 penalties on the parameter difference. Most of our results are for the L_2 penalty. For the L_2 penalty, the resulting estimate is a linear combination of the two within sample estimates. Theorem 1 gives a formula for the degrees of freedom of that estimate. Theorem 2 presents the mean squared error of the estimator and forms the basis for plug-in estimation of an oracle's value when an L_2 penalty is used.

Section 3 considers in detail the case where the regression simplifies to estimation of a population mean. In that setting, we can determine how plug-in, bootstrap and cross-validation estimates of tuning parameters behave. We get an expression for how much information the large sample can add. Theorem 3 gives a soft-thresholding expression for the estimate produced by L_1 penalization and Theorem 4 can be used to find the penalty parameter that an L_1 oracle would choose when the data are Gaussian.

Section 4 presents some simulated examples. We simulate the location problem and find that numerous L_2 penalty methods are admissible, varying in how aggressively they use the larger sample. The L_1 oracle is outperformed by the L_2 oracle in this setting. When the bias is small, the data enrichment methods improve upon the small sample, but when the bias is large then it is best to use the small sample only. Things change when we simulate the regression model. For dimension $d \geqslant 5$, data enrichment outperforms the small sample method in our simulations at all bias levels. We did not see such an inadmissibility outcome when we simulated cases with $d \leqslant 4$.

Section 5 presents our main theoretical result, Theorem 5. When there are 5 or more predictors and 10 or more degrees of freedom for error, then some of our

data enrichment estimators make simply using the small sample inadmissible. The reduction in mean squared error is greatest when the bias is smallest, but no matter how large the bias is, we gain an improvement. This result is similar to Stein's classic result on estimation of a Gaussian mean (Stein, 1956), but the critical threshold here is dimension 5, not dimension 3. The estimator we study employs a data-driven weighting of the two within-sample least squares estimators. We believe that our plug-in estimator is even better than this one.

There are many ideas in different literatures on combining non-identically distributed data sets in order to share or borrow statistical strength. Of these, the closest to our work is small area estimation (Rao, 2003) used in survey sampling. In chemometrics there is a similar problem called transfer calibration (Feudale et al., 2002). Medicine and epidemiology among other fields use meta-analysis (Borenstein et al., 2009). Data fusion (D'Orazio et al., 2006) is widely used in marketing. The problem has been studied for machine learning where it is called transfer learning. An older machine learning term for the underlying issue is concept drift. Bayesian statisticians use hierarchical models. Our methods are more similar to empirical Bayes methods, drawing heavily on ideas of Charles Stein. A Stein-like result also holds for multiple regression in the context of just one sample. The result is intermediate between our two sample regression setting and the one sample mean problem. In regression, shrinkage makes the usual MLE inadmissible when in dimension $p \geqslant 4$ (with the intercept counted as one dimension) and a sufficiently large n. See Copas (1983) for a discussion of shrinkage in regression and Stein (1960) who also obtained this result for regression, but under stronger assumptions.

A more detailed discussion of these different but overlapping literatures is in Section 6. Some of our proofs are given in an (online) Appendix.

There are also settings where one might want to use a small data set to enrich a large one. For example the small data set may have a better design matrix or smaller error variance. Such possibilities are artificial in the motivating context so we don't investigate them further here.

2 Data enrichment regression

Consider linear regression with a response $Y \in \mathbb{R}$ and predictors $X \in \mathbb{R}^d$. The model for the small data set is

$$Y_i = X_i\beta + \varepsilon_i, \quad i \in S$$

for a parameter $\beta \in \mathbb{R}^d$ and independent errors ε_i with mean 0 and variance σ_S^2. Now suppose that the data in the big data set follow

$$Y_i = X_i(\beta + \gamma) + \varepsilon_i, \quad i \in B$$

where $\gamma \in \mathbb{R}^d$ is a bias parameter and ε_i are independent with mean 0 and variance σ_B^2. The sample sizes are n in the small sample and N in the big sample.

There are several kinds of departures of interest. It could be, for instance, that the overall level of Y is different in S than in B but that the trends are similar. That is, perhaps only the intercept component of γ is nonzero. More generally, the effects of some but not all of the components in X may differ in the two samples. One could apply hypothesis testing to each component of γ but that is unattractive as the number of scenarios to test for grows as 2^d.

Let $X_S \in \mathbb{R}^{n\times d}$ and $X_B \in \mathbb{R}^{N\times d}$ have rows made of vectors X_i for $i \in S$ and $i \in B$ respectively. Similarly, let $Y_S \in \mathbb{R}^n$ and $Y_B \in \mathbb{R}^N$ be corresponding vectors of response values. We use $V_S = X_S^{\mathsf{T}} X_S$ and $V_B = X_B^{\mathsf{T}} X_B$.

2.1 Partial pooling via shrinkage and weighting

Our primary approach is to pool the data but put a shrinkage penalty on γ. We estimate β and γ by minimizing

$$\sum_{i\in S}(Y_i - X_i\beta)^2 + \sum_{i\in B}(Y_i - X_i(\beta+\gamma))^2 + \lambda P(\gamma) \tag{1}$$

where $\lambda \in [0,\infty]$ and $P(\gamma) \geqslant 0$ is a penalty function. There are several reasonable choices for the penalty function, including

$$\|\gamma\|_2^2, \quad \|X_S\gamma\|_2^2, \quad \|\gamma\|_1, \quad \text{and} \quad \|X_S\gamma\|_1.$$

For each of these penalties, setting $\lambda = 0$ leads to separate fits $\hat\beta$ and $\hat\beta + \hat\gamma$ in the two data sets. Similarly, taking $\lambda = \infty$ constrains $\hat\gamma = 0$ and amounts to pooling the samples. In many applications one will want to regularize β as well, but in this paper we only penalize γ.

The L_1 penalties have an advantage in interpretation because they identify which parameters or which specific observations might be differentially affected. The quadratic penalties are simpler, so we focus most of this paper on them.

Both quadratic penalties can be expressed as $\|X_T\gamma\|_2^2$ for a matrix X_T. The rows of X_T represent a hypothetical target population of N_T items for prediction. Or more generally, the matrix $\Sigma = \Sigma_T = X_T^{\mathsf{T}} X_T$ is proportional to the matrix of mean squares and mean cross-products for predictors in the target population.

If we want to remove the pooling effect from one of the coefficients, such as the intercept term, then the corresponding column of X_T should contain all zeros. We can also constrain $\gamma_j = 0$ (by dropping its corresponding predictor) in order to enforce exact pooling on the j'th coefficient.

A second, closely related approach is to fit $\hat\beta_S$ by minimizing $\sum_{i\in S}(Y_i - X_i\beta)^2$, fit $\hat\beta_B$ by minimizing $\sum_{i\in B}(Y_i - X_i\beta)^2$, and then estimate β by

$$\hat\beta(\omega) = \omega\hat\beta_S + (1-\omega)\hat\beta_B$$

for some $0 \leqslant \omega \leqslant 1$. In some special cases the estimates indexed by the weighting parameter $\omega \in [n/(n+N), 1]$ are a relabeling of the penalty-based estimates indexed by the parameter $\lambda \in [0,\infty]$. In other cases, the two families of estimates

differ. The weighting approach allows simpler tuning methods. Although we think that the penalization method may be superior, we can prove stronger results about the weighting approach.

Given two values of λ we consider the larger one to be more 'aggressive' in that it makes more use of the big sample bringing with it the risk of more bias in return for a variance reduction. Similarly, aggressive estimators correspond to small weights ω on the small target sample.

2.2 Special cases

An important special case for our applications is the ***cell partition*** model. In the cell partition model, X_i is a vector containing $C-1$ zeros and one 1. The model has C different cells in it. Cell c has N_c observations from the large data set and n_c observations from the small data set. In an advertising context a cell may correspond to one specific demographic subset of consumers. The cells may be chosen exogenously to the given data sets. When the cells are constructed using the regression data then cross-validation or other methods should be used.

A second special case, useful in theoretical investigations, has $X_S^{\mathsf{T}}X_S \propto X_B^{\mathsf{T}}X_B$. This is the ***proportional design matrix*** case.

The simplest case of all is the ***location model***. It is the cell mean model with $C=1$ cell, and it has proportional design matrices. We can get formulas for the optimal tuning parameter in the location model and it is also a good workbench for comparing estimates of tuning parameters. Furthermore, we are able to get some results for the L_1 case in the location model setting.

2.3 Quadratic penalties and degrees of freedom

The quadratic penalty takes the form $P(\gamma) = \|X_T\gamma\|_2^2 = \gamma^{\mathsf{T}}V_T\gamma$ for a matrix $X_T \in \mathbb{R}^{r\times d}$ and $V_T = X_T^{\mathsf{T}}X_T \in \mathbb{R}^{d\times d}$. The value r is d or n in the examples above and could take other values in different contexts. Our criterion becomes

$$\|Y_S - X_S\beta\|^2 + \|Y_B - X_B(\beta+\gamma)\|^2 + \lambda\|X_T\gamma\|^2. \tag{2}$$

Here and below $\|x\|$ means the Euclidean norm $\|x\|_2$.

Given the penalty matrix X_T and a value for λ, the penalized sum of squares (2) is minimized by $\hat\beta_\lambda$ and $\hat\gamma_\lambda$ satisfying

$$\mathcal{X}^{\mathsf{T}}\mathcal{X}\begin{pmatrix}\hat\beta_\lambda\\ \hat\gamma_\lambda\end{pmatrix} = \mathcal{X}^{\mathsf{T}}\mathcal{Y}$$

where

$$\mathcal{X} = \begin{pmatrix} X_{\mathrm{S}} & 0\\ X_{\mathrm{B}} & X_{\mathrm{B}}\\ 0 & \lambda^{1/2}X_T\end{pmatrix} \in \mathbb{R}^{(n+N+r)\times 2d}, \quad \text{and} \quad \mathcal{Y} = \begin{pmatrix} Y_{\mathrm{S}}\\ Y_{\mathrm{B}}\\ 0\end{pmatrix}. \tag{3}$$

To avoid uninteresting complications we suppose that the matrix $\mathcal{X}^{\mathsf{T}}\mathcal{X}$ is invertible. The representation (3) also underlies a convenient computational

approach to fitting $\hat{\beta}_\lambda$ and $\hat{\gamma}_\lambda$ using r rows of pseudo-data just as one does in ridge regression.

The estimate $\hat{\beta}_\lambda$ can be written in terms of $\hat{\beta}_S = V_S^{-1} X_S^\mathsf{T} Y_S$ and $\hat{\beta}_B = V_B^{-1} X_B^\mathsf{T} Y_B$ as the next lemma shows.

Lemma 1. *Let X_S, X_B, and X_T in (2) all have rank d. Then for any $\lambda \geqslant 0$, the minimizers $\hat{\beta}$ and $\hat{\gamma}$ of (2) satisfy*

$$\hat{\beta} = W_\lambda \hat{\beta}_S + (I - W_\lambda)\hat{\beta}_B$$

and $\hat{\gamma} = (V_B + \lambda V_T)^{-1} V_B(\hat{\beta}_B - \hat{\beta})$ for a matrix

$$W_\lambda = (V_S + \lambda V_T V_B^{-1} V_S + \lambda V_T)^{-1}(V_S + \lambda V_T V_B^{-1} V_S). \tag{4}$$

If $V_T = V_S$, then

$$W_\lambda = (V_B + \lambda V_S + \lambda V_B)^{-1}(V_B + \lambda V_S).$$

Proof. The normal equations of (2) are

$$(V_B + V_S)\hat{\beta} = V_S\hat{\beta}_S + V_B\hat{\beta}_B - V_B\hat{\gamma} \quad \text{and} \quad (V_B + \lambda V_T)\hat{\gamma} = V_B\hat{\beta}_B - V_B\hat{\beta}.$$

Solving the second equation for $\hat{\gamma}$, plugging the result into the first and solving for $\hat{\beta}$, yields the result with $W_\lambda = (V_S + V_B - V_B(V_B + \lambda V_T)^{-1} V_B)^{-1} V_S$. This expression for W_λ simplifies as given and simplifies further when $V_T = V_S$. □

The remaining challenge in model fitting is to choose a value of λ. Because we are only interested in making predictions for the S data, not the B data, the ideal value of λ is one that optimizes the prediction error on sample S. One reasonable approach is to use cross-validation by holding out a portion of sample S and predicting the held-out values from a model fit to the held-in ones as well as the entire B sample. One may apply either leave-one-out cross-validation or more general K-fold cross-validation. In the latter case, sample S is split into K nearly equally sized parts and predictions based on sample B and $K-1$ parts of sample S are used for the K'th held-out fold of sample S.

In some of our applications we prefer to use criteria such as AIC, AICc, or BIC in order to avoid the cost and complexity of cross-validation. These alternatives are of most value when data enrichment is itself the inner loop of a more complicated algorithm.

To compute AIC and alternatives, we need to measure the degrees of freedom used in fitting the model. We follow Ye (1998) and Efron (2004) in defining the degrees of freedom to be

$$\mathrm{df}(\lambda) = \frac{1}{\sigma_S^2} \sum_{i \in S} \mathrm{cov}(\hat{Y}_i, Y_i), \tag{5}$$

where $\hat{Y}_S = X_S\hat{\beta}_\lambda$. Because of our focus on the S data, only the S data appear in the degrees of freedom formula. We will see later that the resulting AIC type estimates based on the degrees of freedom perform similarly to our focused cross-validation described above.

Theorem 1. *For data enriched regression the degrees of freedom given at* (5) *satisfies* $\mathrm{df}(\lambda) = \mathrm{tr}(W_\lambda)$ *where* W_λ *is given in Lemma 1. If* $V_T = V_S$*, then*

$$\mathrm{df}(\lambda) = \sum_{j=1}^{d} \frac{1+\lambda\nu_j}{1+\lambda+\lambda\nu_j} \tag{6}$$

where $\nu_1, \ldots, \nu_d$ *are the eigen-values of* $V_S^{1/2} V_B^{-1} V_S^{1/2}$ *in which* $V_S^{1/2}$ *is a symmetric matrix square root of* V_S*.*

Proof. Please see Section 8.1 in the Appendix. □

With a notion of degrees of freedom customized to the data enrichment context we can now define the corresponding criteria such as

$$\begin{aligned} \mathrm{AIC}(\lambda) &= n\log(\hat\sigma_S^2(\lambda)) + n\Big(1 + \frac{2\mathrm{df}(\lambda)}{n}\Big) \quad \text{and} \\ \mathrm{AICc}(\lambda) &= n\log(\hat\sigma_S^2(\lambda)) + n\Big(1 + \frac{\mathrm{df}(\lambda)}{n}\Big) \Big/ \Big(1 - \frac{\mathrm{df}(\lambda)+2}{n}\Big), \end{aligned} \tag{7}$$

where $\hat\sigma_S^2(\lambda) = (n-d)^{-1}\sum_{i\in S}(Y_i - X_i\hat\beta(\lambda))^2$. The AIC is more appropriate than BIC here since our goal is prediction accuracy, not model selection. We prefer the AICc criterion of Hurvich and Tsai (1989) because it is more conservative as the degrees of freedom become large compared to the sample size.

Next we illustrate some special cases of the degrees of freedom formula in Theorem 1. First, suppose that $\lambda = 0$, so that there is no penalization on γ. Then $\mathrm{df}(0) = \mathrm{tr}(I) = d$ as is appropriate for regression on sample S only.

We can easily see that the degrees of freedom are monotone decreasing in λ. As $\lambda \to \infty$ the degrees of freedom drop to $\mathrm{df}(\infty) = \sum_{j=1}^{d} \nu_j/(1+\nu_j)$. This can be much smaller than d. For instance in the proportional design case, $V_S = n\Sigma$ and $V_B = N\Sigma$ for a matrix Σ. Then all $\nu_j = n/N$ and so $\mathrm{df}(\infty) = d/(1+N/n)$, which is quite small when $n \ll N$.

For the cell partition model, d becomes C, $\Sigma_S = \mathrm{diag}(n_c)$ and $\Sigma_B = \mathrm{diag}(N_c)$. In this case $\mathrm{df}(\infty) = \sum_{c=1}^{C} n_c/(n_c+N_c)$ which will usually be much smaller than $\mathrm{df}(0) = C$.

Monotonicity of the degrees of freedom makes it easy to search for the value λ which delivers a desired degrees of freedom. We have found it useful to investigate λ over a numerical grid corresponding to degrees of freedom decreasing from d by an amount Δ (such as 0.25) to the smallest such value above $\mathrm{df}(\infty)$. It is easy to adjoin $\lambda = \infty$ (sample pooling) to this list as well.

2.4 Predictive mean square errors

Here we develop an oracle's choice for λ and a corresponding plug-in estimate. We work in the case where $V_S = V_T$ and we assume that V_S has full rank. Given λ, the predictive mean square error is $\mathbb{E}(\|X_S(\hat\beta - \beta)\|^2)$.

We will use a symmetric square root $V_S^{1/2}$ of V_S as well as the matrix $M = V_S^{1/2}V_B^{-1}V_S^{1/2}$ with eigendecomposition $M = UDU^\mathsf{T}$ where the j'th column of U is u_j and $D = \mathrm{diag}(\nu_j)$.

Theorem 2. *The predictive mean square error of the data enrichment estimator is*

$$\mathbb{E}\big(\|X_S(\hat\beta - \beta)\|^2\big) = \sigma_S^2 \sum_{j=1}^{d} \frac{(1+\lambda\nu_j)^2}{(1+\lambda+\lambda\nu_j)^2} + \sum_{j=1}^{d} \frac{\lambda^2\kappa_j^2}{(1+\lambda+\lambda\nu_j)^2} \tag{8}$$

where $\kappa_j^2 = u_j^\mathsf{T} V_S^{1/2}\Theta V_S^{1/2} u_j$ *for* $\Theta = \gamma\gamma^T + \sigma_B^2 V_B^{-1}$.

Proof. Please see Section 8.2. □

The first term in (8) is a variance term. It equals $d\sigma_S^2$ when $\lambda = 0$ but for $\lambda > 0$ it is reduced due to the use of the big sample. The second term represents the error, both bias squared and variance, introduced by the big sample.

2.5 A plug-in method

A natural choice of λ is to minimize the predictive mean square error, which must be estimated. We propose a plug-in method that replaces the unknown parameters σ_S^2 and κ_j^2 from Theorem 2 by sample estimates. For estimates $\hat\sigma_S^2$ and $\hat\kappa_j^2$ we choose

$$\hat\lambda = \arg\min_{\lambda\geqslant 0} \sum_{j=1}^{d} \frac{\hat\sigma_S^2(1+\lambda\nu_j)^2 + \lambda^2\hat\kappa_j^2}{(1+\lambda+\lambda\nu_j)^2}. \tag{9}$$

From the sample data we take $\hat\sigma_S^2 = \|Y_S - X_S\hat\beta_S\|^2/(n-d)$. A straightforward plug-in estimate of Θ is

$$\widehat\Theta_{\mathrm{plug}} = \hat\gamma\hat\gamma^\mathsf{T} + \hat\sigma_B^2 V_B^{-1},$$

where $\hat\gamma = \hat\beta_B - \hat\beta_S$. Now we take $\hat\kappa_j^2 = u_j^\mathsf{T} V_S^{1/2}\widehat\Theta V_S^{1/2} u_j$ recalling that u_j and ν_j derive from the eigendecomposition of $M = V_S^{1/2}V_B^{-1}V_S^{1/2}$. The resulting optimization yields an estimate $\hat\lambda_{\mathrm{plug}}$.

The estimate $\widehat\Theta_{\mathrm{plug}}$ is biased upwards because $\mathbb{E}(\hat\gamma\hat\gamma^\mathsf{T}) = \gamma\gamma^\mathsf{T} + \sigma_B^2 V_B^{-1} + \sigma_S^2 V_S^{-1}$. We have used a bias-adjusted plug-in estimate

$$\widehat\Theta_{\mathrm{bapi}} = \hat\sigma_B^2 V_B^{-1} + (\hat\gamma\hat\gamma^\mathsf{T} - \hat\sigma_B^2 V_B^{-1} - \hat\sigma_S^2 V_S^{-1})_+ \tag{10}$$

where the positive part operation on a symmetric matrix preserves its eigenvectors but replaces any negative eigenvalues by 0. Similar results can be obtained with $\widetilde\Theta_{\mathrm{bapi}} = \big(\hat\gamma\hat\gamma^\mathsf{T} - \hat\sigma_S^2 V_S^{-1}\big)_+$. This latter estimator is somewhat simpler but the former has the advantage of being at least as large as $\hat\sigma_B^2 V_B^{-1}$ while the latter can degenerate to 0.

3 The location model

The simplest instance of our problem is the location model where X_S is a column of n ones and X_B is a column of N ones. Then the vector β is simply a scalar intercept that we call μ and the vector γ is a scalar mean difference that we call δ. The response values in the small data set are $Y_i = \mu + \varepsilon_i$ while those in the big data set are $Y_i = (\mu + \delta) + \varepsilon_i$. Every quadratic penalty defines the same family of estimators as we get using penalty $\lambda\delta^2$.

The quadratic criterion is $\sum_{i\in S}(Y_i - \mu)^2 + \sum_{i\in B}(Y_i - \mu - \delta)^2 + \lambda\delta^2$. Taking $V_S = n$, $V_B = N$ and $V_T = 1$ in Lemma 1 yields

$$\hat{\mu} = \omega\bar{Y}_S + (1-\omega)\bar{Y}_B \quad \text{with} \quad \omega = \frac{nN + n\lambda}{nN + n\lambda + N\lambda} = \frac{1 + \lambda/N}{1 + \lambda/N + \lambda/n}.$$

Choosing a value for ω corresponds to choosing

$$\lambda = \frac{nN(1-\omega)}{N\omega - n(1-\omega)}.$$

The degrees of freedom in this case reduce to $\mathrm{df}(\lambda) = \omega$, which ranges from $\mathrm{df}(0) = 1$ down to $\mathrm{df}(\infty) = n/(n+N)$.

3.1 Oracle estimator of ω

The mean square error of $\hat{\mu}(\omega)$ is

$$\mathrm{MSE}(\omega) = \omega^2\frac{\sigma_S^2}{n} + (1-\omega)^2\Big(\frac{\sigma_B^2}{N} + \delta^2\Big).$$

The mean square optimal value of ω (available to an oracle) is

$$\omega_{\mathrm{orcl}} = \frac{\delta^2 + \sigma_B^2/N}{\delta^2 + \sigma_B^2/N + \sigma_S^2/n}.$$

Pooling the data corresponds to $\omega_{\mathrm{pool}} = n/(N+n)$ and makes $\hat{\mu}$ equal the pooled mean $\bar{Y}_P \equiv (n\bar{Y}_S + N\bar{Y}_B)/(n+N)$. Ignoring the large data set corresponds to $\omega_S = 1$. Here $\omega_{\mathrm{pool}} \leqslant \omega_{\mathrm{orcl}} \leqslant \omega_S$. The oracle's choice of ω can be used to infer the oracle's choice of λ. It is

$$\lambda_{\mathrm{orcl}} = \frac{nN(1-\omega_{\mathrm{orcl}})}{N\omega_{\mathrm{orcl}} - n(1-\omega_{\mathrm{orcl}})} = \frac{N\sigma_S^2}{N\delta^2 + \sigma_B^2 - \sigma_S^2}. \tag{11}$$

The mean squared error reduction for the oracle is

$$\frac{\mathrm{MSE}(\omega_{\mathrm{orcl}})}{\mathrm{MSE}(\omega_S)} = \omega_{\mathrm{orcl}}, \tag{12}$$

after some algebra. If $\delta \neq 0$, then as $\min(n, N) \to \infty$ we find $\omega_{\mathrm{orcl}} \to 1$ and the optimal λ corresponds to simply using the small sample and ignoring the large

one. If we suppose that $\delta \neq 0$ and $N \to \infty$ then the effective sample size for data enrichment may be defined using (12) as

$$\widetilde{n} = \frac{n}{\omega_{\mathrm{orcl}}} = n\frac{\delta^2 + \sigma_B^2/N + \sigma_S^2/n}{\delta^2 + \sigma_B^2/N} \to n + \frac{\sigma_S^2}{\delta^2}. \tag{13}$$

The mean squared error from data enrichment with n observations in the small sample, using the oracle's choice of λ, matches that of $\widetilde{n}$ IID observations from the small sample. We effectively gain up to σ_S^2/δ^2 observations worth of information. This is an upper bound on the gain because we will have to estimate λ.

Equation (13) shows that the benefit from data enrichment is a small sample phenomenon. The effect is additive not multiplicative on the small sample size n. As a result, more valuable gains are expected in small samples. In some of the motivating examples we have found the most meaningful improvements from data enrichment on disaggregated data sets, such as specific groups of consumers. Some large data sets resemble the union of a great many small ones.

3.2 Plug-in and other estimators of ω

A natural approach to choosing ω is to plug in sample estimates

$$\hat{\delta}_0 = \bar{Y}_B - \bar{Y}_S, \quad \hat{\sigma}_S^2 = \frac{1}{n}\sum_{i\in S}(Y_i - \bar{Y}_S)^2, \quad \text{and} \quad \hat{\sigma}_B^2 = \frac{1}{N}\sum_{i\in B}(Y_i - \bar{Y}_B)^2.$$

We then use $\omega_{\mathrm{plug}} = (\hat{\delta}_0^2 + \hat{\sigma}_B^2/N)/(\hat{\delta}_0^2 + \hat{\sigma}_B^2/N + \hat{\sigma}_S^2/n)$ or alternatively $\lambda_{\mathrm{plug}} = \hat{\sigma}_S^2/(\hat{\sigma}_B^2 + N\hat{\delta}_0^2)$. Our bias-adjusted plug-in method reduces to

$$\omega_{\mathrm{bapi}} = \frac{\hat{\theta}_{\mathrm{bapi}}}{\hat{\theta}_{\mathrm{bapi}} + \hat{\sigma}_S^2/n}, \quad \text{where} \quad \hat{\theta}_{\mathrm{bapi}} = \frac{\hat{\sigma}_B^2}{N} + \Big(\hat{\delta}_0^2 - \frac{\hat{\sigma}_S^2}{n} - \frac{\hat{\sigma}_B^2}{N}\Big)_+$$

The simpler alternative $\widetilde{\omega}_{\mathrm{bapi}} = ((\hat{\delta}_0^2 - \hat{\sigma}_S^2/n)/\hat{\delta}_0^2)_+$ gave virtually identical values in our numerical results reported below.

If we bootstrap the S and B samples independently M times and choose ω to minimize

$$\frac{1}{M}\sum_{m=1}^{M}\big(\bar{Y}_S - \omega\bar{Y}_S^{m*} - (1-\omega)\bar{Y}_B^{m*}\big)^2,$$

then the minimizing value tends to ω_{plug} as $M \to \infty$. Thus bootstrap methods give an approach analogous to plug-in methods, when no simple plug-in formula exists. This is perhaps not surprising since the bootstrap is often described as an example of a plug-in principle.

We can also determine the effects of cross-validation in the location setting, and arrive at an estimate of ω that we can use without actually cross-validating. Consider splitting the small sample into K parts that are held out one by one

in turn. The $K-1$ retained parts are used to estimate μ and then the squared error is judged on the held-out part. That is

$$\omega_{\mathrm{cv}} = \arg\min_{\omega} \frac{1}{K}\sum_{k=1}^{K}\bigl(\bar{Y}_{S,k} - \omega\bar{Y}_{S,-k} - (1-\omega)\bar{Y}_B\bigr)^2,$$

where $\bar{Y}_{S,k}$ is the average of Y_i over the k'th part of S and $\bar{Y}_{S,-k}$ is the average of Y_i over all $K-1$ parts excluding the k'th. We suppose for simplicity that $n = rK$ for an integer r. In that case $\bar{Y}_{S,-k} = (n\bar{Y}_S - r\bar{Y}_{S,k})/(n-r)$. Now

$$\omega_{\mathrm{cv}} = \frac{\sum_k(\bar{Y}_{S,-k} - \bar{Y}_B)(\bar{Y}_{S,k} - \bar{Y}_B)}{\sum_k(\bar{Y}_{S,-k} - \bar{Y}_B)^2} \tag{14}$$

After some algebra, the numerator of (14) is

$$K(\bar{Y}_S - \bar{Y}_B)^2 - \frac{r}{n-r}\sum_{k=1}^{K}(\bar{Y}_{S,k} - \bar{Y}_S)^2$$

and the denominator is

$$K(\bar{Y}_S - \bar{Y}_B)^2 + \left(\frac{r}{n-r}\right)^2\sum_{k=1}^{K}(\bar{Y}_{S,k} - \bar{Y}_S)^2.$$

Letting $\hat{\delta}_0 = \bar{Y}_B - \bar{Y}_S$ and $\hat{\sigma}^2_{S,K} = (1/K)\sum_{k=1}^{K}(\bar{Y}_{S,k} - \bar{Y}_S)^2$, we have

$$\omega_{\mathrm{cv}} = \frac{\hat{\delta}_0^2 - \hat{\sigma}^2_{S,K}/(K-1)}{\hat{\delta}_0^2 + \hat{\sigma}^2_{S,K}/(K-1)^2}.$$

The only quantity in ω_{cv} which depends on the specific K-way partition used is $\hat{\sigma}^2_{S,K}$. If the groupings are chosen by sampling without replacement, then under this sampling,

$$\mathbb{E}(\hat{\sigma}^2_{S,K}) = \mathbb{E}((\bar{Y}_{S,1} - \bar{Y}_S)^2) = \frac{s_S^2}{r}(1 - 1/K)$$

using the finite population correction for simple random sampling, where $s_S^2 = \hat{\sigma}_S^2 n/(n-1)$. This simplifies to

$$\mathbb{E}(\hat{\sigma}^2_{S,K}) = \hat{\sigma}_S^2\frac{n}{n-1}\frac{1}{r}\frac{K-1}{K} = \hat{\sigma}_S^2\frac{K-1}{n-1}.$$

Thus K-fold cross-validation chooses a weighting centered around

$$\omega_{\mathrm{cv},K} = \frac{\hat{\delta}_0^2 - \hat{\sigma}_S^2/(n-1)}{\hat{\delta}_0^2 + \hat{\sigma}_S^2/[(n-1)(K-1)]}. \tag{15}$$

Cross-validation has the strange property that $\omega < 0$ is possible. This can arise when the bias is small and then sampling alone makes the held-out part of the

small sample appear negatively correlated with the held-in part. The effect can appear with any K. We replace any $\omega_{\mathrm{cv},K} < n/(n+N)$ by $n/(n+N)$.

Leave-one-out cross-validation has $K = n$ (and $r = 1$) so that

$$\omega_{\mathrm{cv},n} \approx \frac{\hat{\delta}_0^2 - \hat{\sigma}_S^2/n}{\hat{\delta}_0^2 + \hat{\sigma}_S^2/n^2}.$$

Smaller K, such as choosing $K = 10$ versus n, tend to make $\omega_{\mathrm{cv},K}$ smaller resulting in less weight on $\bar{Y}_S$. In the extreme with $\hat{\delta}_0 = 0$ we find $\omega_{\mathrm{cv},K} \approx -(K-1)$ so 10 fold CV is then very different from leave-one-out CV.

Remark 1. The cross-validation estimates do not make use of $\hat{\sigma}_B^2$ because the large sample is held fixed. They are in this sense conditional on the large sample. Our oracle takes account of the randomness in set B, so it is not conditional. One can define a conditional oracle without difficulty, but we omit the details. Neither the bootstrap nor the plug-in methods are conditional, as they approximate our oracle. Comparing cross-validation to the oracle we expect this to be reasonable if $\sigma_B^2/N \ll \min(\delta^2, \sigma_s^2/n)$. Taking ω_{bapi} as a representor of unconditional methods and $\omega_{\mathrm{cv},n}$ as a representor of conditional ones, we see that the latter has a larger denominator while they both have the same numerator, at least when $\hat{\delta}_0^2 > \hat{\sigma}_S^2/n$. This suggests that conditional methods are more aggressive and we will see this in the simulation results.

3.3 L_1 penalty

For the location model, it is convenient to write the L_1 penalized criterion as

$$\sum_{i\in S}(Y_i - \mu)^2 + \sum_{i\in B}(Y_i - \mu - \delta)^2 + 2\lambda|\delta|. \tag{16}$$

The minimizers $\hat{\mu}$ and $\hat{\delta}$ satisfy

$$\begin{aligned} \hat{\mu} &= \frac{n\bar{Y}_S + N(\bar{Y}_B - \hat{\delta})}{n+N}, \quad \text{and} \\ \hat{\delta} &= \Theta(\bar{Y}_B - \hat{\mu}; \lambda/N) \end{aligned} \tag{17}$$

for the well-known soft thresholding operator $\Theta(z;\tau) = \mathrm{sign}(z)(|z| - \tau)_+$.

The estimate $\hat{\mu}$ ranges from $\bar{Y}_S$ at $\lambda = 0$ to the pooled mean $\bar{Y}_P$ at $\lambda = \infty$. In fact $\hat{\mu}$ reaches $\bar{Y}_P$ at a finite value $\lambda = \lambda_* \equiv nN|\bar{Y}_B - \bar{Y}_S|/(N+n)$ and both $\hat{\mu}$ and $\hat{\delta}$ are linear in λ on the interval $[0, \lambda_*]$:

Theorem 3. *If $0 \leqslant \lambda \leqslant nN|\bar{Y}_B - \bar{Y}_S|/(n+N)$ then the minimizers of* (16) *are*

$$\begin{aligned} \hat{\mu} &= \bar{Y}_S + \frac{\lambda}{n}\,\mathrm{sign}(\bar{Y}_B - \bar{Y}_S), \quad \textit{and} \\ \hat{\delta} &= \bar{Y}_B - \bar{Y}_S - \lambda\frac{N+n}{Nn}\,\mathrm{sign}(\bar{Y}_B - \bar{Y}_S). \end{aligned} \tag{18}$$

If $\lambda > nN|\bar{Y}_B - \bar{Y}_S|/(n+N)$ then they are $\hat{\delta} = 0$ and $\hat{\mu} = \bar{Y}_P$.

Proof. If $\lambda > nN|\bar{Y}_B - \bar{Y}_S|/(n+N)$ then we may find directly that with any value of $\delta > 0$ and corresponding μ given by (17), the derivative of (16) with respect to δ is positive. Therefore $\hat{\delta} \leqslant 0$ and a similar argument gives $\hat{\delta} \geqslant 0$, so that $\hat{\delta} = 0$ and then $\hat{\mu} = (n\bar{Y}_S + N\bar{Y}_B)/(n+N)$.

Now suppose that $\lambda \leqslant \lambda_*$. We verify that the quantities in (18) jointly satisfy equations (17). Substituting $\hat{\delta}$ from (18) into the first line of (17) yields

$$\frac{n\bar{Y}_S + N(\bar{Y}_S + \lambda(N+n)\eta/(Nn))}{n+N} = \bar{Y}_S + \frac{\lambda}{n}\mathrm{sign}(\bar{Y}_B - \bar{Y}_S),$$

matching the value in (18). Conversely, substituting $\hat{\mu}$ from (18) into the second line of (17) yields

$$\Theta\Big(\bar{Y}_B - \hat{\mu}; \frac{\lambda}{N}\Big) = \Theta\Big(\bar{Y}_B - \bar{Y}_S - \frac{\lambda}{n}\mathrm{sign}(\bar{Y}_B - \bar{Y}_S); \frac{\lambda}{N}\Big). \tag{19}$$

Because of the upper bound on λ, the result is $\bar{Y}_B - \bar{Y}_S - \lambda(1/n + 1/N)\mathrm{sign}(\bar{Y}_B - \bar{Y}_S)$ which matches the value in (18). □

With an L_1 penalty on δ we find from Theorem 3 that

$$\hat{\mu} = \bar{Y}_S + \min(\lambda, \lambda_*)\mathrm{sign}(\bar{Y}_B - \bar{Y}_S)/n.$$

That is, the estimator moves $\bar{Y}_S$ towards $\bar{Y}_B$ by an amount λ/n except that it will not move past the pooled average $\bar{Y}_P$. The optimal choice of λ is not available in closed form.

3.4 An L_1 oracle

Under a Gaussian data assumption, it is possible to derive a formula for the mean squared error of the L_1 penalized data enrichment estimator at any value of λ. While it is unwieldy, the L_1 mean square error formula is computable and we can optimize it numerically to compute an oracle formula. As with the L_2 setting we must plug in estimates of some unknowns first before optimizing. This allows us to compare L_1 to L_2 penalization in the location setting simulations of Section 4.

To obtain a solution we make a few changes of notation just for this subsection. We replace λ/n by λ and define $a = N/(N+n)$ and use $\hat{\delta}_0 = \bar{Y}_B - \bar{Y}_S$. Then

$$\begin{aligned}\hat{\mu}(\lambda) &= (\bar{Y}_S + \lambda \cdot \mathrm{sign}(\hat{\delta}_0))I(|\hat{\delta}_0|a \geqslant \lambda) + (a\bar{Y}_B + (1-a)\bar{Y}_S)I(|\hat{\delta}_0|a < \lambda) \\ &= (a\bar{Y}_B + (1-a)\bar{Y}_S) - (a\hat{\delta}_0 - \lambda \cdot \mathrm{sign}(\hat{\delta}_0))I(|\hat{\delta}_0|a \geqslant \lambda). \end{aligned} \tag{20}$$

Without loss of generality we may center and scale the Gaussian distributions so that $\bar{Y}_S \sim \mathcal{N}(0,1)$ and $\bar{Y}_B \sim \mathcal{N}(\delta, \sigma^2)$. The next Theorem defines the distributions of Y_i for $i \in S$ and $i \in B$ to obtain that scaling. We also introduce constants $b = \sigma^2/(1+\sigma^2)$, $\tilde{\delta} = \delta/\sqrt{1+\sigma^2}$, $\tilde{x} = (\lambda/a)/\sqrt{1+\sigma^2}$, and the function $g(x) = \Phi(x) - x\varphi(x)$ where φ and Φ are the $\mathcal{N}(0,1)$ probability density function and cumulative distribution function, respectively.

Theorem 4. *Suppose that* $Y_i \overset{iid}{\sim} \mathcal{N}(0, n)$ *for* $i \in S$ *independently of* $Y_i \overset{iid}{\sim} \mathcal{N}(\delta, \sigma^2 N)$ *for* $i \in B$. *Let* $\hat{\mu}$ *be the* L_1 *estimate from* (20), *using parameter* $\lambda \geqslant 0$. *Then the predictive mean squared error is*

$$\begin{aligned}\mathbb{E}(\hat{\mu}(\lambda)^2) &= a^2\delta^2 + (a+b-1)^2(1+\sigma^2) + b \\ &\quad - a(a+2b-2)(1+\sigma^2)[1 - g(\tilde{x}-\tilde{\delta}) + g(-\tilde{x}-\tilde{\delta})] \\ &\quad - [2a\lambda + 2(a+b-1)(a\delta-\lambda)]\sqrt{1+\sigma^2}\varphi(\tilde{x}-\tilde{\delta}) \\ &\quad - [2a\lambda - 2(a+b-1)(a\delta+\lambda)]\sqrt{1+\sigma^2}\varphi(-\tilde{x}-\tilde{\delta}) \\ &\quad - (a\delta-\lambda)(a\delta+\lambda)[1 - \Phi(\tilde{x}-\tilde{\delta}) + \Phi(-\tilde{x}-\tilde{\delta})].\end{aligned} \tag{21}$$

Proof. Please see Section 8.3 in the Appendix. □

3.5 Cell means

The cell mean setting is simply C copies of the location problem. One could estimate separate values of λ in each of them. Here we remark briefly on the consequences of using a common λ or ω over all cells.

We do not simulate the various choices. We look instead at what assumptions would make them match the oracle formula. In applications we can choose the method whose matching assumptions are more plausible.

In the L_2 setting, one could choose a common λ using either the penalty $\lambda\sum_{c=1}^{C} n_c\delta_c^2$ or $\lambda\sum_{c=1}^{C}\delta_c^2$. Call these cases $L_{2,n}$ and $L_{2,1}$ respectively. Dropping the subscript c we find

$$\omega_{L_{2,n}} = \frac{1+\lambda n/N}{1+\lambda n/N+\lambda}, \quad \text{and} \quad \omega_{L_{2,1}} = \frac{1+\lambda/N}{1+\lambda/N+\lambda/n}$$

compared to $\omega_{\mathrm{orcl}} = (n\delta^2+\sigma_B^2 n/N)/(n\delta^2+\sigma_B^2 n/N+\sigma_S^2)$.

We can find conditions under which a single value of λ recovers the oracle's weighting. For $\omega_{L_{2,1}}$ these are $\sigma_{B,c}^2 = \sigma_{S,c}^2$ in all cells as well as $\lambda = \sigma_{S,c}^2/\delta_c^2$ constant in c. For $\omega_{L_{2,n}}$ these are $\sigma_{B,c}^2 = \sigma_{S,c}^2$ and $\lambda = \sigma_{S,c}^2/(n_c\delta_c^2)$ constant in c. The $L_{2,1}$ criterion looks more reasonable here because we have no reason to expect the relative bias $\delta_c/\sigma_{S,c}$ to be inversely proportional to $\sqrt{n_c}$.

For a common ω to match the oracle, we need $\sigma_{B,c}^2/N_c = \sigma_{S,c}^2/n_c$ to hold in all cells as well as a $\sigma_{S,c}^2/(n_c\delta_c^2)$ to be constant in c. The first clause seems quite unreasonable and so we prefer common-λ approaches to common weights.

For a common L_1 penalty, we cannot get good expressions for the weight variable ω. But we can see how the L_1 approach shifts the mean. An $L_{1,1}$ approach moves $\hat{\mu}_c$ from $\bar{Y}_{S,c}$ towards $\bar{Y}_{B,c}$ by the amount λ/n_c in cell c, but not going past the pooled mean $\bar{Y}_{P,c} = (n\bar{Y}_{S,c} + N\bar{Y}_{B,c})/(N+n)$ for that cell. The other approaches use different shifts. An $L_{1,n}$ approach moves $\hat{\mu}_c$ from $\bar{Y}_{S,c}$ towards $\bar{Y}_{B,c}$ by the amount λ in cell c (but not past $\bar{Y}_{P,c}$). It does not seem reasonable to move $\hat{\mu}_c$ by the same distance in all cells, or to move them by an amount proportional to $1/n_c$ and stopping at $\bar{Y}_{P,c}$ doesn't fix this. We could use a common moving distance proportional to $1/\sqrt{n_c}$ (which is the order of statistical uncertainty in $\bar{Y}_{S,c}$) by using the penalty $\sum_{c=1}^{C}\sqrt{n_c}|\gamma_c|$.

4 Numerical examples

We have simulated some special cases of the data enrichment problem. First we simulate the pure location problem which has $d = 1$. Then we consider the regression problem with varying d.

4.1 Location

We simulated Gaussian data for the location problem. The large sample had $N = 1000$ observations and the small sample had $n = 100$ observations: $X_i \sim \mathcal{N}(\mu, \sigma_S^2)$ for $i \in S$ and $X_i \sim \mathcal{N}(\mu + \delta, \sigma_B^2)$ for $i \in B$. Our data had $\mu = 0$ and $\sigma_S^2 = \sigma_B^2 = 1$. We define the relative bias as

$$\delta_* = \frac{|\delta|}{\sigma_S/\sqrt{n}} = \sqrt{n}|\delta|.$$

We investigated a range of relative bias values. It is only a small simplification to take $\sigma_S^2 = \sigma_B^2$. Doubling σ_B^2 has a very similar effect to halving N. Equal variances might have given a slight relative advantage to the hypothesis testing method as described below.

The accuracy of our estimates is judged by the relative mean squared error $\mathbb{E}((\hat{\mu} - \mu)^2)/(\sigma_S^2/n)$. Simply taking $\hat{\mu} = \bar{Y}_S$ attains a relative mean squared error of 1.

Figure 1 plots relative mean squared error versus relative bias for a collection of estimators, with the results averaged over 10,000 simulated data sets. We used the small sample only method as a control variate.

The solid curve in Figure 1 shows the oracle's value. It lies strictly below the horizontal S-only line. None of the competing curves lie strictly below that line. None can because $\bar{Y}_S$ is an admissible estimator for $d = 1$ (Stein, 1956). The second lowest curve in Figure 1 is for the oracle using the L_1 version of the penalty. The L_1 penalized oracle is not as effective as the L_2 oracle and it is also more difficult to approximate. The highest observed predictive MSEs come from a method of simply pooling the two samples. That method is very successful when the relative bias is near zero but has an MSE that becomes unbounded as the relative bias increases.

Now we discuss methods that use the data to decide whether to use the small sample only, pool the samples or choose an amount of shrinkage. We may list them in order of their worst case performance. From top (worst) to bottom (best) in Figure 1 they are: hypothesis testing, 5-fold cross-validation, 10-fold cross-validation, AICc, leave-one-out cross-validation, and then the simple plug-in method which is minimax among this set of choices. AICc and leave-one-out are very close. Our cross-validation estimators used $\omega = \max(\omega_{\mathrm{cv},K}, n/(n+N))$ where $\omega_{\mathrm{cv},K}$ is given by (15).

The hypothesis testing method is based on a two-sample t-test of whether $\delta = 0$. If the test is rejected at $\alpha = 0.05$, then only the small sample data is used. If the test is not rejected, then the two samples are pooled. That test was

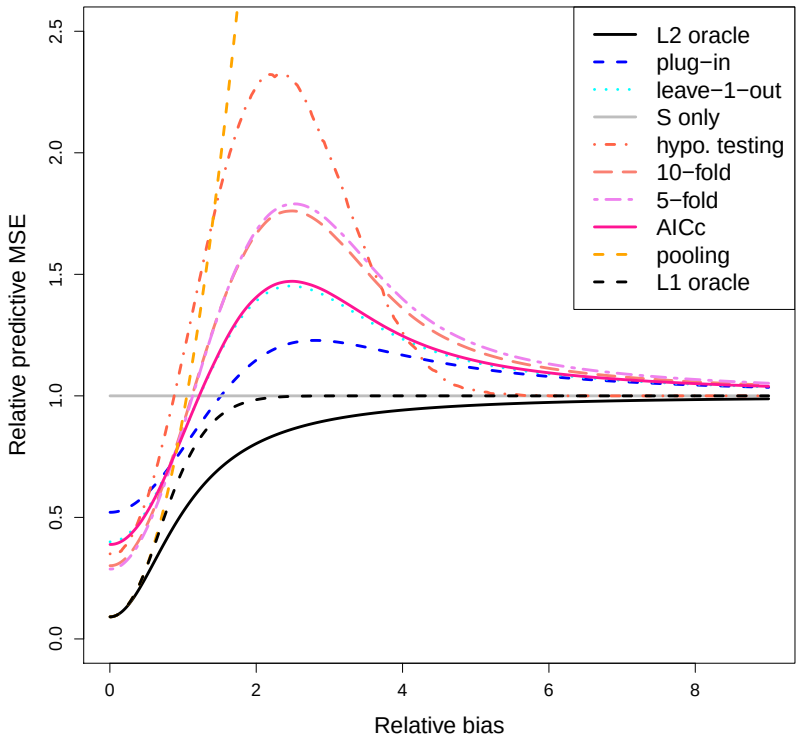

Figure 1: Numerical results for the location problem. The horizontal line at 1 represents using the small sample only and ignoring the large one. The lowest line shown is for an oracle choosing λ in the L_2 penalization. The green curve shows an oracle using the L_1 penalization. The other curves are as described in the text.

based on $\sigma_B^2 = \sigma_S^2$ which may give hypothesis testing a slight advantage in this setting (but it still performed poorly).

The AICc method performs virtually identically to leave-one-out cross-validation over the whole range of relative biases.

None of these methods makes any other one inadmissible: each pair of curves crosses. The methods that do best at large relative biases tend to do worst at relative bias near 0 and vice versa. The exception is hypothesis testing. Compared to the others it does not benefit fully from low relative bias but it recovers the quickest as the bias increases. Of these methods hypothesis testing is best at the highest relative bias, K-fold cross-validation with small K is best at the lowest relative bias, and the plug-in method is best in between.

Aggressive methods will do better at low bias but worse at high bias. What

we see in this simulation is that K-fold cross-validation is the most aggressive followed by leave-one-out and AICc and that plug-in is least aggressive. These findings confirm what we saw in the formulas from Section 3. Hypothesis testing does not quite fit into this spectrum: its worst case performance is much worse than the most aggressive methods yet it fails to fully benefit from pooling when the bias is smallest. Unlike aggressive methods it does very well at high bias.

4.2 Regression

We simulated our data enrichment method for the following scenario. The small sample had $n = 1000$ observations and the large sample had $N = 10{,}000$. The true β was taken to be 0. This has no loss of generality because we are not shrinking β towards 0. The value of γ was taken uniformly on the unit sphere in d dimensions and then multiplied by a scale factor that we varied.

We considered $d = 2, 4, 5$ and 10. All of our examples included an intercept column of 1s in both X_S and X_B. The other $d-1$ predictors were sampled from a Gaussian distribution with covariance C_S or C_B, respectively. In one simulation we took C_S and C_B to be independent Wishart$(I, d-1, d-1)$ random matrices. In the other they were sampled as $C_S = I_{d-1} + \rho uu^{\mathsf{T}}$ and $C_B = I_{d-1} + \rho vv^{\mathsf{T}}$ where u and v are independently and uniformly sampled from the unit sphere in $\mathbb{R}^{d-1}$ and $\rho \geqslant 0$ is a parameter that measures the lack of proportionality between covariances. We chose $\rho = d$ so that the sample specific portion of the variance has comparable magnitude to the common part.

We scaled the results so that regression using sample S only yields a mean squared error of 1 at all values of the relative bias. We computed the risk of an L_2 oracle, as well as sampling errors when λ is estimated by the plug-in formula, by our bias-adjusted plug-in formula and via AICc. In addition we considered the simple weighted combination $\omega\hat{\beta}_S + (1-\omega)\hat{\beta}_B$ with ω chosen by the plug-in formula.

Figure 2 shows the results. For $d = 2$ and also $d = 4$ none of our methods universally outperforms simply using the S sample. For $d = 5$ and $d = 10$, all of our estimators have lower mean squared error than using the S sample alone, though the difference becomes small at large relative bias.

We find in this setting that our bias-adjusted plug-in estimator closely matches the AICc estimate. The relative performance of the other methods varies with the problem. Plain plug-in always seemed worse than AICc and adjusted plug-in at low relative bias and better than these at high biases. Plug-in's gains at high biases appear to be less substantial than its losses at low biases. Of the other methods, simple scalar weighting is worst for the high dimensional Wishart case without being better in the other cases. The best overall choices are bias-adjusted plug-in and AICc.

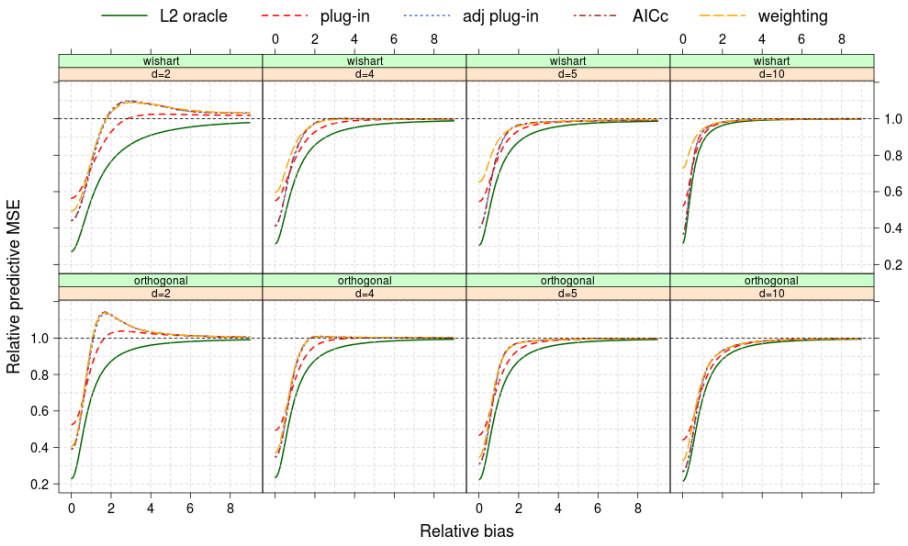

Figure 2: This figure shows relative predicted MSE versus relative bias for two simulated regression problems described in the text.

5 Proportional design and inadmissibility

The proportional design case has $V_B \propto V_S$ and $V_T \propto V_S$. Suppose that $V_B = N\Sigma$, $V_S = n\Sigma$ and $V_T = \Sigma$ for a positive definite matrix Σ. Our data enrichment estimator simplifies greatly in this case. The weighting matrix W_λ in Lemma 1 simplifies to $W_\lambda = \omega I$ where $\omega = (N + n\lambda)/(N + n\lambda + N\lambda)$. As a result $\hat{\beta} = \omega\hat{\beta}_S + (1-\omega)\hat{\beta}_B$ and we can find and estimate an oracle's value for ω. If different constants of proportionality, say M and m are used, then the effect is largely to reparameterize λ giving the same family of estimates under different labels. There is one difference though. The interval of possible values for ω is $[n/N, 1]$ in our case versus $[m/M, 1]$ for the different constants. To attain the same sets of ω values could require use of negative λ.

The resulting estimator of $\hat{\beta}$ with estimated ω dominates $\hat{\beta}_S$ (making it inadmissible) under mild conditions. These conditions given below even allow violations of the proportionality condition $V_B \propto V_S$ but they still require $V_T \propto V_S$. Among these conditions we will need the model degrees of freedom to be at least 5, and it will suffice to have the error degrees of freedom in the small sample regression be at least 10. The result also requires a Gaussian assumption in order to use a lemma of Stein's.

We write $Y_S = X_S\beta + \varepsilon_S$ and $Y_B = X_B(\beta + \gamma) + \varepsilon_B$ for $\varepsilon_S \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_S^2)$ and $\varepsilon_B \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_B^2)$. The data enrichment estimators are $\hat{\beta}(\lambda)$ and $\hat{\gamma}(\lambda)$. The parameter of most interest is β. If we were to use only the small sample we would get $\hat{\beta}_S = (X_S^\mathsf{T} X_S)^{-1} X_S^\mathsf{T} Y_S = \hat{\beta}(0)$.

In the proportional design setting, the mean squared prediction error is

$$\begin{aligned} f(\omega) &= \mathbb{E}(\|X_T(\hat{\beta}(\omega) - \beta)\|^2) \\ &= \mathrm{tr}((\omega^2\sigma_S^2\Sigma_S^{-1} + (1-\omega)^2(\gamma\gamma^\mathsf{T} + \sigma_B^2\Sigma_B^{-1}))\Sigma). \end{aligned}$$

This error is minimized by the oracle's parameter value

$$\omega_{\mathrm{orcl}} = \frac{\mathrm{tr}((\gamma\gamma^\mathsf{T} + \sigma_B^2\Sigma_B^{-1})\Sigma)}{\mathrm{tr}((\gamma\gamma^\mathsf{T} + \sigma_B^2\Sigma_B^{-1})\Sigma) + \sigma_S^2\mathrm{tr}(\Sigma_S^{-1}\Sigma)}.$$

With $\Sigma_S = n\Sigma$ and $\Sigma_B = N\Sigma$, we find

$$\omega_{\mathrm{orcl}} = \frac{\gamma^\mathsf{T}\Sigma\gamma + d\sigma_B^2/N}{\gamma^\mathsf{T}\Sigma\gamma + d\sigma_B^2/N + d\sigma_S^2/n}.$$

The plug-in estimator is

$$\hat{\omega}_{\mathrm{plug}} = \frac{\hat{\gamma}^\mathsf{T}\Sigma\hat{\gamma} + d\hat{\sigma}_B^2/N}{\hat{\gamma}^\mathsf{T}\Sigma\hat{\gamma} + d\hat{\sigma}_B^2/N + d\hat{\sigma}_S^2/n} \tag{22}$$

where $\hat{\sigma}_S^2 = \|Y_S - X_S\hat{\beta}_S\|^2/(n-d)$ and $\hat{\sigma}_B^2 = \|Y_B - X_B\hat{\beta}_B\|^2/(N-d)$. We will have reason to generalize this plug-in estimator. Let $h(\hat{\sigma}_B^2)$ be any nonnegative measurable function of $\hat{\sigma}_B^2$ with $\mathbb{E}(h(\hat{\sigma}_B^2)) < \infty$. The generalized plug-in estimator is

$$\hat{\omega}_{\mathrm{plug},h} = \frac{\hat{\gamma}^\mathsf{T}\Sigma\hat{\gamma} + h(\hat{\sigma}_B^2)}{\hat{\gamma}^\mathsf{T}\Sigma\hat{\gamma} + h(\hat{\sigma}_B^2) + d\hat{\sigma}_S^2/n}. \tag{23}$$

Here are the conditions under which $\hat{\beta}_S$ is made inadmissible by the data enrichment estimator.

Theorem 5. *Let $X_S \in \mathbb{R}^{n\times d}$ and $X_B \in \mathbb{R}^{N\times d}$ be fixed matrices with $X_S^\mathsf{T}X_S = n\Sigma$ and $X_B^\mathsf{T}X_B = N\Sigma_B$ where Σ and Σ_B both have rank d. Let $Y_S \sim \mathcal{N}(X_S\beta, \sigma_S^2 I_n)$ independently of $Y_B \sim \mathcal{N}(X_B(\beta+\gamma), \sigma_B^2 I_N)$. If $d \geqslant 5$ and $m \equiv n - d \geqslant 10$, then*

$$\mathbb{E}(\|X_T\hat{\beta}(\hat{\omega}) - X_T\beta\|^2) < \mathbb{E}(\|X_T\hat{\beta}_S - X_T\beta\|^2) \tag{24}$$

holds for any nonrandom matrix X_T with $X_T^\mathsf{T}X_T = \Sigma$ and any $\hat{\omega} = \hat{\omega}_{\mathrm{plug},h}$ given by (23).

Proof. Please see Section 8.5 in the Appendix. □

The condition on m can be relaxed at the expense of a more complicated statement. From the details in the proof, it suffices to have $d \geqslant 5$ and $m(1 - 4/d) \geqslant 2$.

The result in Theorem 5 is similar to the Stein estimator result. There, the sample mean of a Gaussian population is an inadmissible estimator in $d = 3$

dimensions or higher but is admissible in 1 or 2 dimensions. Here there are two samples to pool and the change takes place at $d = 5$.

Because $\mathbb{E}(\hat\gamma^{\mathsf{T}}\Sigma\hat\gamma) = \gamma^{\mathsf{T}}\Sigma\gamma + d\sigma_S^2/n + d\sigma_B^2/N$ it is biased high and so therefore is $\hat\omega_{\mathrm{plug}}$, making it a little conservative. We can make a bias adjustment, replacing $\hat\gamma^{\mathsf{T}}\Sigma\hat\gamma$ by $\hat\gamma^{\mathsf{T}}\Sigma\hat\gamma - d\hat\sigma_S^2/n - d\hat\sigma_B^2/N$. The result is

$$\hat\omega_{\mathrm{bapi}} = \frac{\hat\gamma^{\mathsf{T}}\Sigma\hat\gamma - d\hat\sigma_S^2/n}{\hat\gamma^{\mathsf{T}}\Sigma\hat\gamma} \vee \frac{n}{n+N}, \tag{25}$$

where values below $n/(n+N)$ get rounded up. This bias-adjusted estimate of ω is not covered by Theorem 5. Subtracting only $\hat\sigma_B^2/N$ instead of $\hat\sigma_B^2/N + \hat\sigma_S^2/n$ is covered, yielding

$$\hat\omega'_{\mathrm{bapi}} = \frac{\hat\gamma^{\mathsf{T}}\Sigma\hat\gamma}{\hat\gamma^{\mathsf{T}}\Sigma\hat\gamma + d\hat\sigma_S^2/n}, \tag{26}$$

which corresponds to taking $h(\hat\sigma_B^2) \equiv 0$ in equation (23).

6 Related literatures

There are many disjoint literatures that study problems like the one we have presented. They do not seem to have been compared before and the literatures seem to be mostly unaware of each other. We give a summary of them here, kept brief because of space limitations.

The key ingredient in this problem is that we care more about the small sample than the large one. Were that not the case, we could simply pool all the data and fit a model with indicator variables picking out one or indeed many different small areas. Without some kind of regularization, that approach ends up being similar to taking $\lambda = 0$ and hence does not borrow strength.

The closest match to our problem setting comes from small area estimation in survey sampling. The monograph by Rao (2003) is a comprehensive treatment of that work and Ghosh and Rao (1994) provide a compact summary. In that context the large sample may be census data from the entire country and the small sample (called the small area) may be a single county or a demographically defined subset. Every county or demographic group may be taken to be the small sample in its turn. The composite estimator (Rao, 2003, Chapter 4.3) is a weighted sum of estimators from small and large samples. The estimates being combined may be more complicated than regressions, involving for example ratio estimates. The emphasis is usually on scalar quantities such as small area means or totals, instead of the regression coefficients we consider. One particularly useful model (Ghosh and Rao, 1994, Equation (4.2)) allows the small areas to share regression coefficients apart from an area specific intercept. Then BLUP estimation methods lead to shrinkage estimators similar to ours.

The methods of Copas (1983) can be applied to our problem and will result in another combination that makes $\hat\beta_S$ inadmissible. That combination requires

only four dimensional regressions instead of the five used in Theorem 5 for pooling weights. That combination yields less aggressive predictions.

In chemometrics a calibration transfer problem (Feudale et al., 2002) comes up when one wants to adjust a model to new spectral hardware. There may be a regression model linking near-infrared spectroscopy data to a property of some sample material. The transfer problem comes up for data from a new machine. Sometimes one can simply run a selection of samples through both machines but in other cases that is not possible, perhaps because one machine is remote (Woody et al., 2004). Their primary and secondary instruments correspond to our small and big samples respectively. Their emphasis is on transfering either principal components regression or partial least squares models, not the plain regressions we consider here.

A common problem in marketing is data fusion, also known as statistical matching. Variables (X, Y) are measured in one sample while variables (X, Z) are measured in another. There may or may not be a third sample with some measured triples (X, Y, Z). The goal in data fusion is to use all of the data to form a large synthetic data set of (X, Y, Z) values, perhaps by imputing missing Z for the (X, Y) sample and/or missing Y for the (X, Z) sample. When there is no (X, Y, Z) sample some untestable assumptions must be made about the joint distribution, because it cannot be recovered from its bivariate margins. The text by D'Orazio et al. (2006) gives a comprehensive summary of what can and cannot be done. Many of the approaches are based on methods for handling missing data (Little and Rubin, 2009).

Our problem is an instance of what machine learning researchers call domain adaptation. They may have fit a model to a large data set (the 'source') and then wish to adapt that model to a smaller specialized data set (the 'target'). This is especially common in natural language processing. NIPS 2011 included a special session on domain adaptation. In their motivating problems there are typically a very large number of features (e.g., one per unique word appearing in a set of documents). They also pay special attention to problems where many of the data points do not have a measured response. Quite often a computer can gather high dimensional X while a human rater is necessary to produce Y. Daumé (2009) surveys various wrapper strategies, such as fitting a model to weighted combinations of the data sets, deriving features from the reference data set to use in the target one and so on. Cortes and Mohri (2011) consider domain adaptation for kernel-based regularization algorithms, including kernel ridge regression, support vector machines (SVMs), or support vector regression (SVR). They prove pointwise loss guarantees depending on the discrepancy distance between the empirical source and target distributions, and demonstrate the power of the approach on a number of experiments using kernel ridge regression.

A related term in machine learning is concept drift (Widmer and Kubat, 1996). There a prediction method may become out of date as time goes on. The term drift suggests that slow continual changes are anticipated, but they also consider that there may be hidden contexts (latent variables in statistical teminology) affecting some of the data.

7 Conclusions

We have studied a middle ground between pooling a large data set into a smaller target one and ignoring it completely. In dimension $d \geqslant 5$ only a small number of error degrees of freedom suffice to make ignoring the large data set inadmissible. When there is no bias, pooling the data sets may be optimal. Theorem 5 does not say that pooling is inadmissible. When there is no bias, pooling the data sets may be optimal. We prefer our hybrid because the risk from pooling grows without bound as the bias increases.

Acknowledgments

We thank the following people for helpful discussions: Penny Chu, Corinna Cortes, Tony Fagan, Yijia Feng, Jerome Friedman, Jim Koehler, Diane Lambert, Elissa Lee and Nicolas Remy.

References

Borenstein, M., Hedges, L. V., Higgins, J. P. T., and Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. Wiley, Chichester, UK.

Copas, J. B. (1983). Regression, prediction and shrinkage. *Journal of the Royal Statistical Society, Series B*, 45(3):311–354.

Cortes, C. and Mohri, M. (2011). Domain adaptation in regression. In *Proceedings of The 22nd International Conference on Algorithmic Learning Theory (ALT 2011)*, pages 308–323, Heidelberg, Germany. Springer.

Daumé, H. (2009). Frustratingly easy domain adaptation. (arXiv:0907.1815).

D'Orazio, M., Di Zio, M., and Scanu, M. (2006). *Statistical Matching: Theory and Practice*. Wiley, Chichester, UK.

Efron, B. (2004). The estimation of prediction error. *Journal of the American Statistical Association*, 99(467):619–632.

Feudale, R. N., Woody, N. A., Tan, H., Myles, A. J., Brown, S. D., and Ferré, J. (2002). Transfer of multivariate calibration models: a review. *Chemometrics and Intelligent Laboratory Systems*, 64:181–192.

Ghosh, M. and Rao, J. N. K. (1994). Small area estimation: an appraisal. *Statistical Science*, 9(1):55–76.

Hurvich, C. and Tsai, C. (1989). Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307.

Little, R. J. A. and Rubin, D. B. (2009). *Statistical Analysis with Missing Data*. John Wiley & Sons Inc., Hoboken, NJ, 2nd edition.

Rao, J. N. K. (2003). *Small Area Estimation*. Wiley, Hoboken, NJ.

Stein, C. M. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley symposium on mathematical statistics and probability*, volume 1, pages 197–206.

Stein, C. M. (1960). Multiple regression. In Olkin, I., Ghurye, S. G., Hoeffding, W., Madow, W. G., and Mann, H. B., editors, *Contributions to probability and statistics: essays in honor of Harald Hotelling*. Stanford University Press, Stanford, CA.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151.

Widmer, G. and Kubat, M. (1996). Learning in the presence of concept drift and hidden contexts. *Machine Learning*, 23:69–101.

Woody, N. A., Feudale, R. N., Myles, A. J., and Brown, S. D. (2004). Transfer of multivariate calibrations between four near-infrared spectrometers using orthogonal signal correction. *Analytical Chemistry*, 76(9):2596–2600.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93:120–131.

8 Appendix: proofs

This appendix presents proofs of the results in this article. They are grouped into sections by topic, with some technical supporting lemmas separated into their own sections.

8.1 Proof of Theorem 1

First $\mathrm{df}(\lambda) = \sigma_S^{-2}\mathrm{tr}(\mathrm{cov}(X_S\hat{\beta}, Y_S)) = \sigma_S^{-2}\mathrm{tr}(X_S W_\lambda (X_S^\mathsf{T} X_S)^{-1} X_S^\mathsf{T}\sigma_S^2) = \mathrm{tr}(W_\lambda)$. Next with $X_T = X_S$, and $M = V_S^{1/2}V_B^{-1}V_S^{1/2}$,

$$\mathrm{tr}(W_\lambda) = \mathrm{tr}(V_S + \lambda V_S V_B^{-1} V_S + \lambda V_S)^{-1}(V_S + \lambda V_S V_B^{-1} V_S).$$

We place $V_S^{1/2}V_S^{-1/2}$ between these factors and absorb them left and right. Then we reverse the order of the factors and repeat the process, yielding

$$\mathrm{tr}(W_\lambda) = \mathrm{tr}(I + \lambda M + \lambda I)^{-1}(I + \lambda M).$$

Writing $M = U\mathrm{diag}(\nu_1, \ldots, \nu_d)U^\mathsf{T}$ for an orthogonal matrix U and simplifying yields the result. □

8.2 Proof of Theorem 2

Proof. First $\mathbb{E}(\|X_T\hat{\beta} - X_T\beta\|^2) = \mathrm{tr}(V_S\mathbb{E}((\hat{\beta}-\beta)(\hat{\beta}-\beta)^\mathsf{T}))$. Next using $W = W_\lambda$, we make a bias-variance decomposition,

$$\begin{aligned}\mathbb{E}\big((\hat{\beta}-\beta)(\hat{\beta}-\beta)^\mathsf{T}\big) &= (I-W)\gamma\gamma^\mathsf{T}(I-W)^\mathsf{T} + \mathrm{cov}(W\hat{\beta}_S) + \mathrm{cov}((I-W)\hat{\beta}_B)\\ &= \sigma_S^2 W V_S^{-1} W^\mathsf{T} + (I-W)\Theta(I-W)^\mathsf{T},\end{aligned}$$

for $\Theta = \gamma\gamma^\mathsf{T} + \sigma_B^2 V_B^{-1}$. Therefore $\mathbb{E}\big(\|X_S(\hat{\beta}-\beta)\|^2\big) = \sigma_S^2\mathrm{tr}(V_S W V_S^{-1} W^\mathsf{T}) + \mathrm{tr}(\Theta(I-W)^\mathsf{T} V_S (I-W))$.

Now we introduce $\widetilde{W} = V_S^{1/2} W V_S^{-1/2}$ finding

$$\begin{aligned}\widetilde{W} &= V_S^{1/2}(V_B + \lambda V_S + \lambda V_B)^{-1}(V_B + \lambda V_S)V_S^{-1/2}\\ &= (I + \lambda M + \lambda I)^{-1}(I + \lambda M)\\ &= U\widetilde{D}U^\mathsf{T},\end{aligned}$$

where $\widetilde{D} = \mathrm{diag}((1+\lambda\nu_j)/(1+\lambda+\lambda\nu_j))$. This allows us to write the first term of the mean squared error as

$$\sigma_S^2\mathrm{tr}(V_S W V_S^{-1} W^\mathsf{T}) = \sigma_S^2\mathrm{tr}(\widetilde{W}\widetilde{W}^\mathsf{T}) = \sigma_S^2\sum_{j=1}^d \frac{(1+\lambda\nu_j)^2}{(1+\lambda+\lambda v_j)^2}.$$

For the second term, let $\widetilde{\Theta} = V_S^{1/2}\Theta V_S^{1/2}$. Then

$$\begin{aligned}\mathrm{tr}\big(\Theta(I-W)^\mathsf{T}V_S(I-W)\big) &= \mathrm{tr}(\widetilde{\Theta}(I-\widetilde{W})^\mathsf{T}(I-\widetilde{W}))\\ &= \mathrm{tr}(\tilde{\Theta}U(I-\widetilde{D})^2U^\mathsf{T})\\ &= \lambda^2\sum_{k=1}^d \frac{u_k^\mathsf{T}V_S^{1/2}\Theta V_S^{1/2}u_k}{(1+\lambda+\lambda\nu_k)^2}.\end{aligned}$$

□

8.3 Proof of Theorem 4

We will use this small lemma.

Lemma 2. *If $X \sim \mathcal{N}(0,1)$, then $\mathbb{E}(XI(X \leqslant x)) = -\varphi(x)$, $\mathbb{E}(X^2 I(X \leqslant x)) = g(x)$ and*

$$\mathbb{E}(X^2 I(|X+b| \geqslant x)) = 1 - g(x-b) + g(-x-b)$$

where $g(x) = \Phi(x) - x\varphi(x)$.

Proof. First $\mathbb{E}(XI(X \leqslant x)) = \int_{-\infty}^x z\varphi(z)\,\mathrm{d}z = -\int_{-\infty}^x \varphi'(z)\,\mathrm{d}z = -\varphi(x)$. Next,

$$\int_{-\infty}^x z^2\varphi(z)\,\mathrm{d}z = -\int_{-\infty}^x z\varphi'(z)\,\mathrm{d}z = -\int_{-\infty}^x \varphi(z)\,\mathrm{d}z - z\varphi(z)\Big|_{-\infty}^x = g(x).$$

Then

$$\begin{aligned}\mathbb{E}(X^2 I(|X+b| \geqslant x)) &= \mathbb{E}(X^2 I(X+b \geqslant x)) + \mathbb{E}(X^2 I(X+b \leqslant -x)) \\ &= \mathbb{E}(X^2(1 - I(X+b \leqslant x)) + g(-x-b)) \\ &= \mathbb{E}(X^2) - \mathbb{E}(X^2 I(X+b \leqslant x)) + g(-x-b) \\ &= 1 - g(x-b) + g(-x-b). \end{aligned}$$ □

Now we prove Theorem 4. We let $\epsilon = \hat{\delta}_0 - \delta$ and $\eta = \bar{Y}_B + \sigma^2 \bar{Y}_S - \delta$. Then

$$\operatorname{cov}(\epsilon, \eta) = 0, \quad \epsilon \sim \mathcal{N}(0, 1+\sigma^2), \quad \eta \sim \mathcal{N}(0, \sigma^2 + \sigma^4), \quad \text{and} \quad \bar{Y}_S = \frac{\eta - \epsilon}{1+\sigma^2}.$$

Recall that we defined $b = \sigma^2/(1+\sigma^2)$, and so

$$\bar{Y}_B = \delta + \eta - \sigma^2 \frac{\eta - \epsilon}{1+\sigma^2} = \delta + b\epsilon + (1-b)\eta.$$

Also with $a = N/(N+n)$,

$$\begin{aligned} a\bar{Y}_B + (1-a)\bar{Y}_S &= a\delta + a(b\epsilon + (1-b)\eta) + (1-a)\frac{\eta - \epsilon}{1+\sigma^2} \\ &= a\delta + (ab - (1-a)(1-b))\epsilon + (a(1-b) + (1-a)(1-b))\eta \\ &= a\delta + (a+b-1)\epsilon + (1-b)\eta. \end{aligned}$$

Letting $S = \epsilon + \delta$, we have

$$\hat{\mu} = a\delta + (a+b-1)\epsilon + (1-b)\eta - (aS - \lambda \cdot \operatorname{sign}(S)) I(|S| \geqslant a^{-1}\lambda)$$

from which the MSE can be calculated:

$$\begin{aligned} \mathbb{E}(\hat{\mu}^2(\lambda)) &= \mathbb{E}\big((a\delta + (a+b-1)\epsilon + (1-b)\eta)^2\big) \\ &\quad - 2\mathbb{E}\big((a\delta + (a+b-1)\epsilon + (1-b)\eta)(aS - \lambda \cdot \operatorname{sign}(S)) I(|S| \geqslant a^{-1}\lambda)\big) \\ &\quad + \mathbb{E}\big((aS - \lambda \cdot \operatorname{sign}(S))^2 I(|S| \geqslant a^{-1}\lambda)\big) \\ &\equiv [1] - 2 \times [2] + [3]. \end{aligned}$$

First

$$\begin{aligned}[1] &= a^2\delta^2 + (a+b-1)^2(1+\sigma^2) + (1-b)^2\sigma^2(1+\sigma^2) \\ &= a^2\delta^2 + (a+b-1)^2(1+\sigma^2) + b. \end{aligned}$$

Next using $\bar{\Phi}(x) = 1 - \Phi(x)$,

$$
\begin{aligned}
[2] &= \mathbb{E}\big([a\delta + (a+b-1)\epsilon][a(S) - \lambda \cdot \mathrm{sign}(S)]I(|S| \geqslant a^{-1}\lambda)\big) \\
&= \mathbb{E}\big(\{a\delta(a\delta - \lambda \cdot \mathrm{sign}(S)) + \epsilon\,[a\delta a + (a+b-1)(a\delta - \lambda \cdot \mathrm{sign}(S))] + a(a+b-1)\epsilon^2\}I(|S| \geqslant a^{-1}\lambda)\big) \\
&= \mathbb{E}\big(a\delta(a\delta - \lambda \cdot \mathrm{sign}(S))I(|S| \geqslant a^{-1}\lambda)\big) \\
&\quad + \mathbb{E}\big(\big[a^2\delta + (a+b-1)(a\delta - \lambda \cdot \mathrm{sign}(S))\big]\,\epsilon I(|S| \geqslant a^{-1}\lambda)\big) \\
&\quad + \mathbb{E}\big(a(a+b-1)\epsilon^2 I(|S| \geqslant a^{-1}\lambda)\big) \\
&= a\delta(a\delta - \lambda)\bar{\Phi}\left(\frac{a^{-1}\lambda - \delta}{\sqrt{1+\sigma^2}}\right) + a\delta(a\delta + \lambda)\Phi\left(\frac{-a^{-1}\lambda - \delta}{\sqrt{1+\sigma^2}}\right) \\
&\quad + [a^2\delta + (a+b-1)(a\delta - \lambda)]\mathbb{E}\big(\epsilon I(S \geqslant a^{-1}\lambda)\big) \\
&\quad + [a^2\delta + (a+b-1)(a\delta + \lambda)]\mathbb{E}\big(\epsilon I(S < -a^{-1}\lambda)\big) \\
&\quad + a(a+b-1)\mathbb{E}\big(\epsilon^2 I(|S| \geqslant a^{-1}\lambda)\big).
\end{aligned}
$$

Recall that we defined $\tilde{x} = a^{-1}\lambda/\sqrt{1+\sigma^2}$ and $\tilde{\delta} = \delta/\sqrt{1+\sigma^2}$. Now using Lemma 2

$$
\begin{aligned}
\mathbb{E}\big(\epsilon^2 I(|S| \geqslant a^{-1}\lambda)\big) &= (1+\sigma^2)\mathbb{E}\left(X^2 I\Big[|X + \frac{\delta}{\sqrt{(1+\sigma^2)}}| \geqslant \frac{a^{-1}\lambda}{\sqrt{(1+\sigma^2)}}\Big]\right) \\
&= (1+\sigma^2)[1 - g(\tilde{x} - \tilde{\delta}) + g(-\tilde{x} - \tilde{\delta})].
\end{aligned}
$$

Next

$$
\begin{aligned}
\mathbb{E}\big(\epsilon I(|S| \geqslant a^{-1}\lambda)\big) &= \mathbb{E}\big(\epsilon I(S \geqslant a^{-1}\lambda)\big) + \mathbb{E}\big(\epsilon I(S \leqslant -a^{-1}\lambda)\big) \\
&= -\mathbb{E}\big(\epsilon I(S \leqslant a^{-1}\lambda)\big) + \mathbb{E}\big(\epsilon I(S \leqslant -a^{-1}\lambda)\big) \\
&= \sqrt{1+\sigma^2}\varphi(\tilde{x} - \tilde{\delta}) - \sqrt{1+\sigma^2}\varphi(-\tilde{x} - \tilde{\delta}).
\end{aligned}
$$

So,

$$
\begin{aligned}
[2] &= a\delta(a\delta - \lambda)\bar{\Phi}(\tilde{x} - \tilde{\delta}) + a\delta(a\delta + \lambda)\Phi(-\tilde{x} - \tilde{\delta}) \\
&\quad + [a^2\delta + (a+b-1)(a\delta - \lambda)]\sqrt{1+\sigma^2}\varphi(\tilde{x} - \tilde{\delta}) \\
&\quad - [a^2\delta + (a+b-1)(a\delta + \lambda)]\sqrt{1+\sigma^2}\varphi(-\tilde{x} - \tilde{\delta}) \\
&\quad + a(a+b-1)(1+\sigma^2)[1 - g(\tilde{x} - \tilde{\delta}) + g(-\tilde{x} - \tilde{\delta})].
\end{aligned}
$$

Finally,

$$\begin{aligned}
[3] &= \mathbb{E}\big([a(S) - \lambda \cdot \mathrm{sign}(S)]^2 I(|S| \geqslant a^{-1}\lambda)\big) \\
&= \mathbb{E}\big([a^2\epsilon^2 + 2a(a\delta - \lambda \cdot \mathrm{sign}(S))\epsilon + (a\delta - \lambda \cdot \mathrm{sign}(S))^2] I(|S| \geqslant a^{-1}\lambda)\big) \\
&= \mathbb{E}\big(a^2\epsilon^2 I(|S| \geqslant a^{-1}\lambda)\big) \\
&\quad + 2\mathbb{E}\big(a(a\delta - \lambda \cdot \mathrm{sign}(S))\epsilon I(|S| \geqslant a^{-1}\lambda)\big) \\
&\quad + \mathbb{E}\big((a\delta - \lambda \cdot \mathrm{sign}(S))^2 I(|S| \geqslant a^{-1}\lambda)\big) \\
&= a^2(1+\sigma^2)[1 - g(\tilde{x} - \tilde{\delta}) + g(-\tilde{x} - \tilde{\delta})] \\
&\quad + 2a(a\delta - \lambda)\sqrt{1+\sigma^2}\varphi(\tilde{x} - \tilde{\delta}) - 2a(a\delta + \lambda)\sqrt{1+\sigma^2}\varphi(-\tilde{x} - \tilde{\delta}) \\
&\quad + (a\delta - \lambda)^2 \bar{\Phi}(\tilde{x} - \tilde{\delta}) + (a\delta + \lambda)^2 \Phi(-\tilde{x} - \tilde{\delta}).
\end{aligned}$$

Hence, the MSE is

$$\begin{aligned}
\mathbb{E}\big(\hat{\mu}^2\big) &= [1] - 2 \times [2] + [3] \\
&= a^2\delta^2 + (a + b - 1)^2(1 + \sigma^2) + b \\
&\quad - a(a + 2b - 2)(1 + \sigma^2)[1 - g(\tilde{x} - \tilde{\delta}) + g(-\tilde{x} - \tilde{\delta})] \\
&\quad - [2a\lambda + 2(a + b - 1)(a\delta - \lambda)]\sqrt{1+\sigma^2}\varphi(\tilde{x} - \tilde{\delta}) \\
&\quad - [2a\lambda - 2(a + b - 1)(a\delta + \lambda)]\sqrt{1+\sigma^2}\varphi(-\tilde{x} - \tilde{\delta}) \\
&\quad - (a\delta - \lambda)(a\delta + \lambda)[1 - \Phi(\tilde{x} - \tilde{\delta}) + \Phi(-\tilde{x} - \tilde{\delta})]. \ \square
\end{aligned}$$

8.4 Supporting lemmas for inadmissibility

In this section we first recall Stein's Lemma. Then we prove two technical lemmas used in the proof of Theorem 5.

Lemma 3. *Let $Z \sim \mathcal{N}(0,1)$ and let $g : \mathbb{R} \to \mathbb{R}$ be an indefinite integral of the Lebesgue measurable function g', essentially the derivative of g. If $\mathbb{E}(|g'(Z)|) < \infty$ then*

$$\mathbb{E}(g'(Z)) = \mathbb{E}(Zg(Z)).$$

Proof. Stein (1981). $\square$

Lemma 4. *Let $\eta \sim \mathcal{N}(0, I_d)$, $b \in \mathbb{R}^d$, and let $A > 0$ and $B > 0$ be constants. Let*

$$Z = \eta + \frac{A(b - \eta)}{\|b - \eta\|^2 + B}.$$

Then

$$\begin{aligned}
\mathbb{E}(\|Z\|^2) &= d + \mathbb{E}\left(\frac{A(A + 4 - 2d)}{\|b - \eta\|^2 + B}\right) - \mathbb{E}\left(\frac{AB(A+4)}{(\|b - \eta\|^2 + B)^2}\right) \\
&< d + \mathbb{E}\left(\frac{A(A + 4 - 2d)}{\|b - \eta\|^2 + B}\right).
\end{aligned}$$

Proof. First,

$$\mathbb{E}(\|Z\|^2) = d + \mathbb{E}\left(\frac{A^2\|b-\eta\|^2}{(\|b-\eta\|^2+B)^2}\right) + 2A\sum_{k=1}^{d}\mathbb{E}\left(\frac{\eta_k(b_k-\eta_k)}{\|b-\eta\|^2+B}\right).$$

Now define

$$g(\eta_k) = \frac{b_k-\eta_k}{\|b-\eta\|^2+B} = \frac{b_k-\eta_k}{(b_k-\eta_k)^2+\|b_{-k}-\eta_{-k}\|^2+B}.$$

By Stein's lemma (Lemma 3), we have

$$\mathbb{E}\left(\frac{\eta_k(b_k-\eta_k)}{\|b-\eta\|^2+B}\right) = \mathbb{E}(g'(\eta_k)) = \mathbb{E}\left(\frac{2(b_k-\eta_k)^2}{(\|b-\eta\|^2+B)^2} - \frac{1}{\|b-\eta\|^2+B}\right)$$

and thus

$$\begin{aligned}
\mathbb{E}(\|Z\|^2) &= d + \mathbb{E}\left(\frac{(4A+A^2)\|b-\eta\|^2}{(\|b-\eta\|^2+B)^2} - \frac{2Ad}{\|b-\eta\|^2+B}\right)\\
&= d + \mathbb{E}\left(\frac{(4A+A^2)\|b-\eta\|^2}{(\|b-\eta\|^2+B)^2} - \frac{2Ad(\|b-\eta\|^2+B)}{(\|b-\eta\|^2+B)^2}\right)\\
&= d + \mathbb{E}\left(\frac{(4A+A^2-2Ad)}{\|b-\eta\|^2+B} - \frac{(4A+A^2)B}{(\|b-\eta\|^2+B)^2}\right),
\end{aligned}$$

after collecting terms. □

Lemma 5. *For integer $m \geqslant 1$, let $Q \sim \chi^2_{(m)}$, $C > 1$, $D > 0$ and put*

$$Z = \frac{Q(C - m^{-1}Q)}{Q+D}.$$

Then

$$\mathbb{E}(Z) \geqslant \frac{(C-1)m-2}{m+2+D}.$$

and so $\mathbb{E}(Z) > 0$ whenever $C > 1 + 2/m$.

Proof. The $\chi^2_{(m)}$ density function is $p_m(x) = (2^{m/2-1}\Gamma(\frac{m}{2}))^{-1}x^{m/2-1}e^{-x/2}$. Thus

$$\begin{aligned}
\mathbb{E}(Z) &= \frac{1}{2^{m/2}\Gamma(\frac{m}{2})}\int_0^\infty \frac{x(C-m^{-1}x)}{x+D}x^{m/2-1}e^{-x/2}\,\mathrm{d}x\\
&= \frac{1}{2^{m/2}\Gamma(\frac{m}{2})}\int_0^\infty \frac{C-m^{-1}x}{x+D}x^{(m+2)/2-1}e^{-x/2}\,\mathrm{d}x\\
&= \frac{2^{m/2+1}\Gamma(\frac{m+2}{2})}{2^{m/2}\Gamma(\frac{m}{2})}\int_0^\infty \frac{C-m^{-1}x}{x+D}p_{m+2}(x)\,\mathrm{d}x\\
&= m\int_0^\infty \frac{C-m^{-1}x}{x+D}p_{m+2}(x)\,\mathrm{d}x\\
&\geqslant m\frac{C-(m+2)/m}{m+2+D}
\end{aligned}$$

by Jensen's inequality. □

8.5 Proof of Theorem 5.

We prove this first for $\hat{\omega}_{\mathrm{plug},h} = \hat{\omega}_{\mathrm{plug}}$, that is, taking $h(\hat{\sigma}_B^2) = d\hat{\sigma}_B^2/n$. We also assume at first that $\Sigma_B = \Sigma$.

Note that $\hat{\beta}_S = \beta + (X_S^\mathsf{T} X_S)^{-1} X_S^\mathsf{T} \varepsilon_S$ and $\hat{\beta}_B = \beta + \gamma + (X_B^\mathsf{T} X_B)^{-1} X_B^\mathsf{T} \varepsilon_B$. It is convenient to define

$$\eta_S = \Sigma^{1/2}(X_S^\mathsf{T} X_S)^{-1} X_S^\mathsf{T} \varepsilon_S \quad \text{and} \quad \eta_B = \Sigma^{1/2}(X_B^\mathsf{T} X_B)^{-1} X_B^\mathsf{T} \varepsilon_B.$$

Then we can rewrite $\hat{\beta}_S = \beta + \Sigma^{-1/2}\eta_S$ and $\hat{\beta}_B = \beta + \gamma + \Sigma^{-1/2}\eta_B$. Similarly, we let

$$\hat{\sigma}_S^2 = \frac{\|Y_S - X_S\hat{\beta}_S\|^2}{n-d} \quad \text{and} \quad \hat{\sigma}_B^2 = \frac{\|Y_B - X_B\hat{\beta}_B\|^2}{N-d}.$$

Now $(\eta_S, \eta_B, \hat{\sigma}_S^2, \hat{\sigma}_B^2)$ are mutually independent, with

$$\begin{aligned} \eta_S &\sim \mathcal{N}\Big(0, \frac{\sigma_S^2}{n} I_d\Big), & \eta_B &\sim \mathcal{N}\Big(0, \frac{\sigma_B^2}{N} I_d\Big), \\ \hat{\sigma}_S^2 &\sim \frac{\sigma_S^2}{n-d}\chi^2_{(n-d)}, \quad \text{and} & \hat{\sigma}_B^2 &\sim \frac{\sigma_B^2}{N-d}\chi^2_{(N-d)}. \end{aligned}$$

We easily find that $\mathbb{E}(\|X\hat{\beta}_S - X\beta\|^2) = d\sigma_S^2/n$. Next we find $\hat{\omega}$ and a bound on $\mathbb{E}(\|X\hat{\beta}(\hat{\omega}) - X\beta\|^2)$.

Let $\gamma^* = \Sigma^{1/2}\gamma$ so that $\hat{\gamma} = \hat{\beta}_B - \hat{\beta}_S = \Sigma^{-1/2}(\gamma^* + \eta_B - \eta_S)$. Then

$$\begin{aligned} \hat{\omega} = \hat{\omega}_{\mathrm{plug}} &= \frac{\hat{\gamma}^\mathsf{T}\Sigma\hat{\gamma} + d\hat{\sigma}_B^2/N}{\hat{\gamma}^\mathsf{T}\Sigma\hat{\gamma} + d\hat{\sigma}_B^2/N + d\hat{\sigma}_S^2/n} \\ &= \frac{\|\gamma^* + \eta_B - \eta_S\|^2 + d\hat{\sigma}_B^2/N}{\|\gamma^* + \eta_B - \eta_S\|^2 + d(\hat{\sigma}_B^2/N + \hat{\sigma}_S^2/n)}. \end{aligned}$$

Now we can express the mean squared error as

$$\begin{aligned} \mathbb{E}(\|X\hat{\beta}(\hat{\omega}) - X\beta\|^2) &= \mathbb{E}(\|X\Sigma^{-1/2}(\hat{\omega}\eta_S + (1-\hat{\omega})(\gamma^* + \eta_B))\|^2) \\ &= \mathbb{E}(\|\hat{\omega}\eta_S + (1-\hat{\omega})(\gamma^* + \eta_B)\|^2) \\ &= \mathbb{E}(\|\eta_S + (1-\hat{\omega})(\gamma^* + \eta_B - \eta_S)\|^2) \\ &= \mathbb{E}\bigg(\Big\|\eta_S + \frac{(\gamma^* + \eta_B - \eta_S)d\hat{\sigma}_S^2/n}{\|\gamma^* + \eta_B - \eta_S\|^2 + d(\hat{\sigma}_B^2/N + \hat{\sigma}_S^2/n)}\Big\|^2\bigg). \end{aligned}$$

To simplify the expression for mean squared error we introduce

$$\begin{aligned} Q &= m\hat{\sigma}_S^2/\sigma_S^2 \sim \chi^2_{(m)} \\ \eta_S^* &= \sqrt{n}\,\eta_S/\sigma_S \sim \mathcal{N}(0, I_d), \\ b &= \sqrt{n}(\gamma^* + \eta_B)/\sigma_S, \\ A &= d\hat{\sigma}_S^2/\sigma_S^2 = dQ/m, \quad \text{and} \\ B &= nd(\hat{\sigma}_B^2/N + \hat{\sigma}_S^2/n)/\sigma_S^2 \\ &= d((n/N)\hat{\sigma}_B^2/\sigma_S^2 + Q/m). \end{aligned}$$

The quantities A and B are, after conditioning, the constants that appear in technical Lemma 4. Similarly C and D introduced below match the constants used in Lemma 5.

With these substitutions and some algebra,

$$\begin{aligned}\mathbb{E}(\|X\hat{\beta}(\hat{\omega}) - X\beta\|^2) &= \frac{\sigma_S^2}{n}\,\mathbb{E}\left(\left\|\eta_S^* + \frac{A(b-\eta_S^*)}{\|b-\eta_S^*\|^2 + B}\right\|^2\right) \\ &= \frac{\sigma_S^2}{n}\,\mathbb{E}\left(\mathbb{E}\left(\left\|\eta_S^* + \frac{A(b-\eta_S^*)}{\|b-\eta_S^*\|^2 + B}\right\|^2 \;\middle|\; \eta_B, \hat{\sigma}_S^2, \hat{\sigma}_B^2\right)\right).\end{aligned}$$

We now apply two technical lemmas from Section 8.4.

Since η_S^* is independent of (b, A, B) and $Q \sim \chi^2_{(m)}$, by Lemma 4, we have

$$\mathbb{E}\left(\left\|\eta_S^* + \frac{A(b-\eta_S^*)}{\|b-\eta_S^*\|^2 + B}\right\|^2 \;\middle|\; \eta_B, \hat{\sigma}_S^2, \hat{\sigma}_B^2\right) < d + \mathbb{E}\left(\frac{A(A+4-2d)}{\|b-\eta_S^*\|^2 + B} \;\middle|\; \eta_B, \hat{\sigma}_S^2, \hat{\sigma}_B^2\right).$$

Hence

$$\begin{aligned}\Delta &\equiv \mathbb{E}(\|X\hat{\beta}_S - X\beta\|^2) - \mathbb{E}(\|X\hat{\beta}(\hat{\omega}) - X\beta\|^2) \\ &= \frac{\sigma_S^2}{n}\mathbb{E}\left(\frac{A(2d-A-4)}{\|b-\eta_S^*\|^2 + B}\right) \\ &= \frac{\sigma_S^2}{n}\mathbb{E}\left(\frac{(dQ/m)(2d - dQ/m - 4)}{\|b-\eta_S^*\|^2 + (B-A) + dQ/m)}\right) \\ &= \frac{d\sigma_S^2}{n}\mathbb{E}\left(\frac{Q(2 - Q/m - 4/d)}{\|b-\eta_S^*\|^2 m/d + (B-A)m/d + Q}\right) \\ &= \frac{d\sigma_S^2}{n}\mathbb{E}\left(\frac{Q(C - Q/m)}{Q + D}\right)\end{aligned}$$

where $C = 2 - 4/d$ and $D = (m/d)(\|b-\eta_S^*\|^2 + dnN^{-1}\hat{\sigma}_B^2/\sigma_S^2)$.

Now suppose that $d \geqslant 5$. Then $C \geqslant 2 - 4/5 > 1$ and so conditionally on η_S, η_B, and $\hat{\sigma}_B^2$, the requirements of Lemma 5 are satisfied by C, D and Q. Therefore

$$\Delta \geqslant \frac{d\sigma_S^2}{n}\mathbb{E}\left(\frac{m(1-4/d) - 2}{m + 2 + D}\right) \tag{27}$$

where the randomness in (27) is only through D which depends on η_S^*, η_B (through b) and $\hat{\sigma}_B^2$. By Jensen's inequality

$$\Delta > \frac{d\sigma_S^2}{n}\frac{m(1-4/d) - 2}{m + 2 + \mathbb{E}(D)} \geqslant 0 \tag{28}$$

whenever $m(1-4/d) \geqslant 2$. The first inequality in (28) is strict because $\mathrm{var}(D) > 0$. Therefore $\Delta > 0$. The condition on m and d holds for any $m \geqslant 10$ when $d \geqslant 5$.

For the general plug-in $\hat{\omega}_{\mathrm{plug},h}$ we replace $d\hat{\sigma}_B^2/N$ above by $h(\hat{\sigma}_B^2)$. This quantity depends on $\hat{\sigma}_B^2$ and is independent of $\hat{\sigma}_S^2$, η_B and η_S. It appears within B where we need it to be non-negative in order to apply Lemma 4. It also appears within D which becomes $(m/d)(\|b-\eta_S^*\|^2 + nh(\hat{\sigma}_B^2)/\sigma_S^2)$. Even when we take $\mathrm{var}(h(\hat{\sigma}_B^2)) = 0$ we still get $\mathrm{var}(D) > 0$ and so the first inequality in (28) is still strict.

Now suppose that Σ_B is not equal to Σ. The distributions of η_S, $\hat{\sigma}_S^2$ and $\hat{\sigma}_B^2$ remain unchanged but now

$$\eta_B \sim \mathcal{N}\Big(0, \frac{\sigma_B^2}{N}\Sigma^{1/2}\Sigma_B^{-1}\Sigma^{1/2}\Big)$$

independently of the others. The changed distribution of η_B does not affect the application of Lemma 4 because that lemma is invoked conditionally on η_B. Similarly, Lemma 5 is applied conditionally on η_B. The changed distribution of η_B changes the distribution of D but we can still apply (28). $\square$